Comparative genomics tools for biological discovery

Inna Dubchak, Ph.D.

Berkeley PGA, Bioinformatics Group Leader

Lawrence Berkeley National Laboratory

ildubchak@lbl.gov

http:/www-gsd.lbl.gov

Outline

What is comparative genomics?

VISTA tools developed for comparative genomics.

Related biological stories

Large scale VISTA applications including automatic computational system for comparing whole vertebrate genomes

The Human genome - 2001

Science
THE HUMAN GENOME
American Association for the Advancement of Science
nature
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special

From the Nature paper:

The next steps:

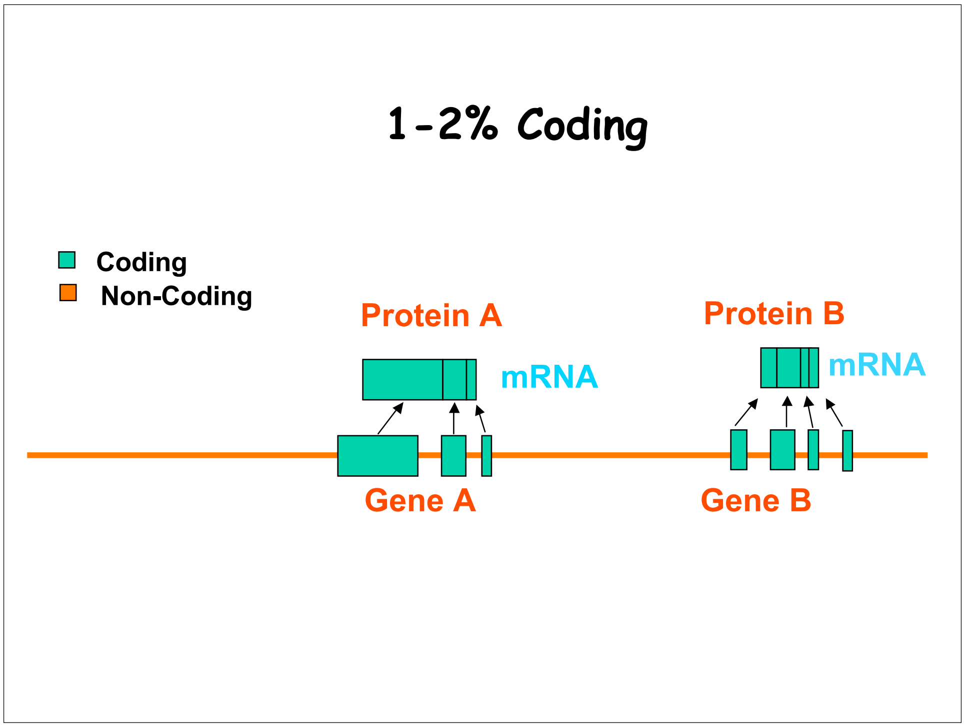
Developing the IGI (integrated gene index) and IPI (integrated protein index)

Large-scale identification of regulatory regions

Sequencing of additional large genomes

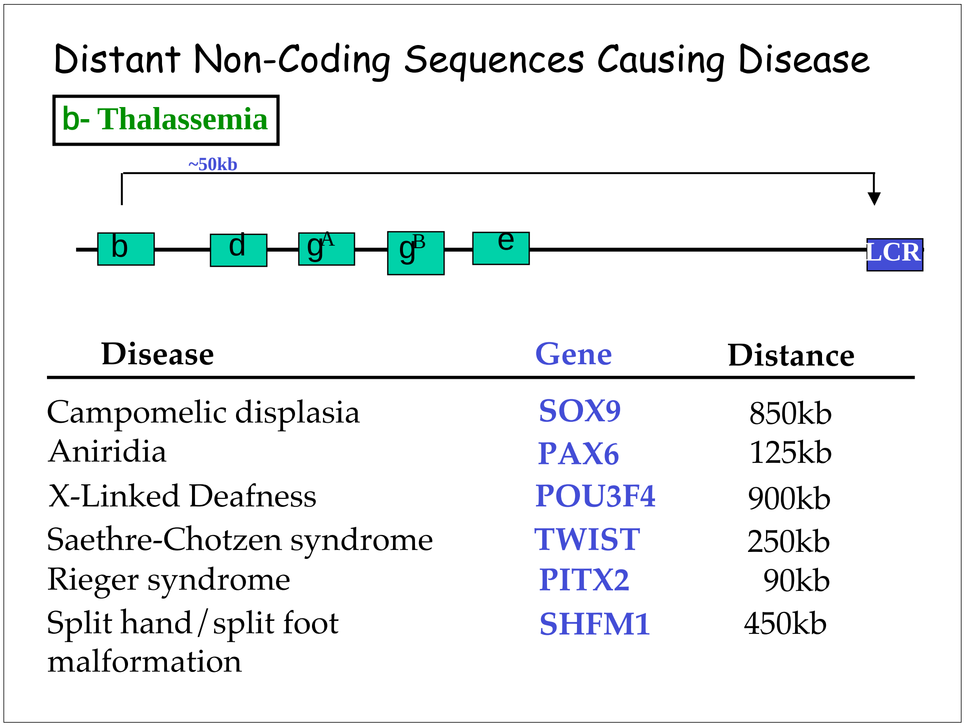
Completing the catalogue of human variation "

From sequence to function



1-2% Coding

🟩 Coding
🟧 Non-Coding



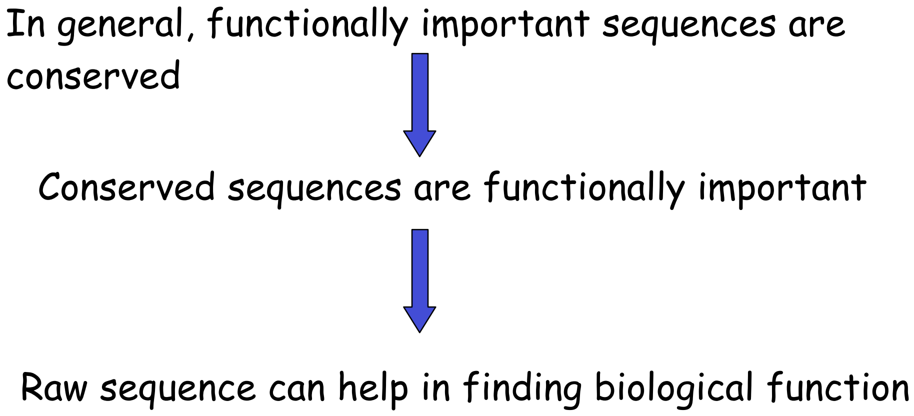
Distant Non-Coding Sequences Causing Disease

b- Thalassemia

Disease	Gene	Distance
Campomelic displasia	SOX9	850kb
Aniridia	PAX6	125kb
X-Linked Deafness	POU3F4	900kb
Saethre-Chotzen syndrome	TWIST	250kb
Rieger syndrome	PITX2	90kb
Split hand/split foot malformation	SHFM1	450kb

Background

Evolution can help!



In general, functionally important sequences are conserved

Conserved sequences are functionally important

Raw sequence can help in finding biological function

Comparison of 1196 orthologous genes (Makalowski et al., 1996)

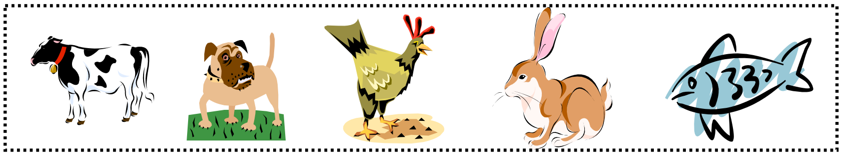
- Sequence identity:

– exons:	84.6%
– protein:	85.4%
– introns:	35%
– 5′ UTRs:	67%
– 3′ UTRs:	69%

- 27 proteins were 100% identical

Integrating data into more powerful gene prediction models than with human genomic sequence alone

Comparing sequences of different organisms



- Helps in gene predictions
- Helps in understanding evolution
- Conserved between species non-coding sequences are reliable guides to regulatory elements
- Differences between evolutionary closely related sequences help to discover gene functions

Sequence comparisons. How?

Three variations:

Find the best OVERALL alignment.

Global alignment

Find ALL regions of similarity.

Local alignment

Find the BEST region of similarity.

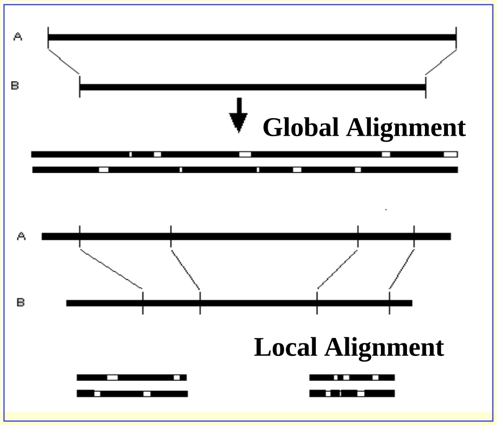
Optimal local alignment

Local alignment algorithms are designed to search for highly similar regions in two sequences that may not be highly similar in their entirety. The algorithm works by first finding very short common segments between the input sequence and database sequences, and then expanding out the matching regions as far as possible.

For cross-species comparison one needs to accurately align two complete sequences. It is insufficient to find common similar regions in the two sequences, rather, what is needed is a global map specifying how the two sequences fit together, much like understanding how the pieces in a puzzle connect up with each other.

This problem is called **global alignment**

Local vs global alignment



Challenges in aligning long genomic regions

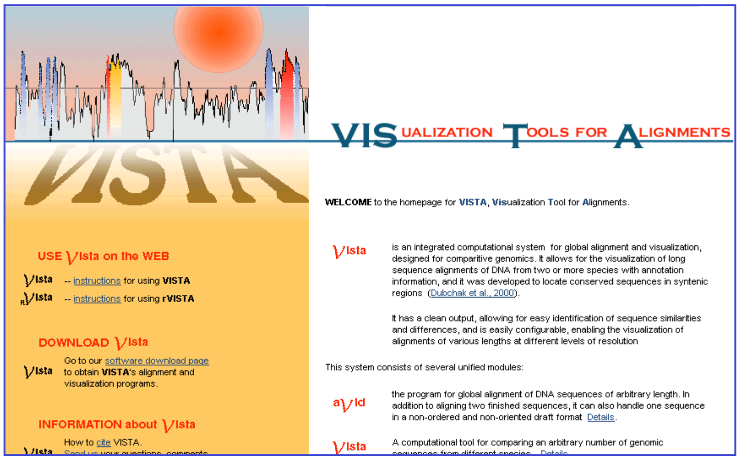
- Long sequences lead to memory problems
- Speed becomes an issue
- Long alignments are very sensitive to parameters
- Draft sequences present a nontrivial problem
- Accuracy is difficult to measure and to achieve
- Scaling up to the size of whole genomes
- Sequence at different stages of completion, difficult to compare

→ Partial Assemblies

Whole genome shotgun

Finished BACs

http://www-gsd.lbl.gov/vista



VISUALIZATION TOOLS FOR ALIGNMENTS

USE Vista on the WEB

Vista -- instructions for using VISTA

RVista -- instructions for using rVISTA

DOWNLOAD Vista

Vista Go to our software download page to obtain VISTA's alignment and visualization programs.

INFORMATION about Vista

Vista How to cite VISTA.
Send us your questions, comments

WELCOME to the homepage for VISTA, Visualization Tool for Alignments.

Vista is an integrated computational system for global alignment and visualization, designed for comparitive genomics. It allows for the visualization of long sequence alignments of DNA from two or more species with annotation information, and it was developed to locate conserved sequences in syntenic regions (Dubchak et al., 2000).

It has a clean output, allowing for easy identification of sequence similarities and differences, and is easily configurable, enabling the visualization of alignments of various lengths at different levels of resolution

This system consists of several unified modules:

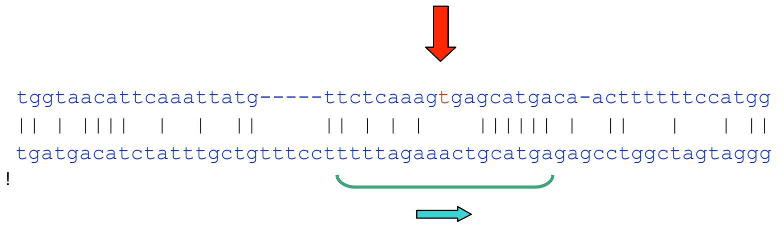
aVId the program for global alignment of DNA sequences of arbitrary length. In addition to aligning two finished sequences, it can also handle one sequence in a non-ordered and non-oriented draft format Details.

Vista A computational tool for comparing an arbitrary number of genomic sequences from different species. Details

Modules of VISTA:

- Program for global alignment of DNA fragments of any length (AVID)
- Visualization of alignment and various sequence features for any number of species
- Evaluation and retrieval of all regions with predefined levels of conservation

Visualization



```
 tggtaacattcaaattatg-----ttctcaaagtgagcatgaca-acttttttccatgg
 || | ||||  |  |  ||     || | | |    |||||| |  ||   |   | ||
 tgatgacatctatttgctgtttcctttttagaaactgcatgagagcctggctagtaggg
!
```

Window of length **L** is centered at a particular nucleotide in the base sequence

Percent of identical nucleotides in **L** positions of the alignment is calculated and plotted

Move to the next nucleotide
!

Finding conserved regions with percentage and length cutoffs

Conserved segments with percent identity X and length Y - regions in which every contiguous subsegment of length Y was at least X% identical to its paired sequence. These segments are merged to define the conserved regions.

Output:

```
11054 - 11156 = 103bp at 77.670%     NONCODING
13241 - 13453 = 213bp at 87.793%      EXON
14698 - 14822 = 125bp at 84.800%      EXON
```

VISTA input files

Sequences

```
> Human ST7 gene
CTGAATGGCTCGTAGAAA
TATTGCATTAACCTGCTG
GACATGCTGAATAGCAAT
CGACTACAGT. .
```

```
> Cow ST7 gene
CTGAATGGCTCGTAGAAA
TAATGCATTCCCCTGCTG
GACATGCTGAATAGCAAT
CGACTACAGT. . . .
```

```
............
```

Annotation for a base sequence if available

```
> 12877 289557 ST7b/a
+ 13076 282515
12877 13226
159297  159379
179096  179255
189328  189382
```

VISTA output files

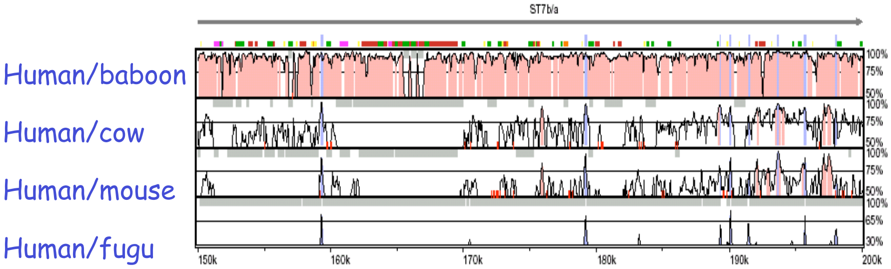
All pair wise alignments

```
185140    185150    185160               185170    185180
GACATTGGAAAAGTAAAGGAAGTGGTTTAT---CTTGCTC------TTTTTGCAACAGTA
         |||| |||||||| | ||||||||||||||    | ||| |      |||||  ||||||
GACACTGGAAAAGCAGAGGAAGTGGTTTATTGACCTGCCCCCCCCTTTTTTATAACAGTG
```

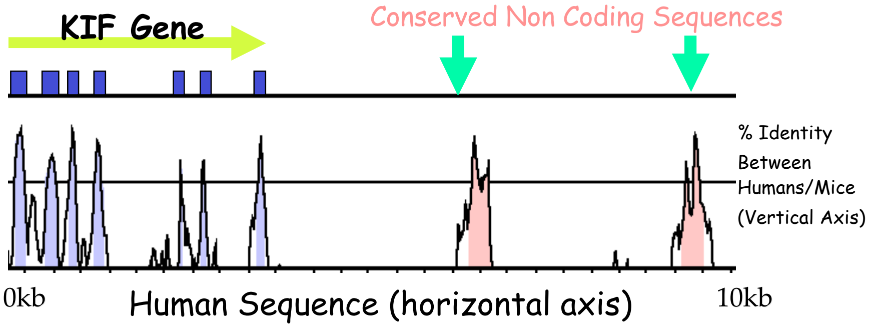
The lists of conserved regions

```
80078 (149626) to  80171 (149724) =   99bp at 63.6% noncoding
 159297 (158141) to  159379 (158223) =    83bp at 80.7% exon
 179096 (159067) to  179253 (159224) =   158bp at 75.9% exon
 189328 (159566) to  189382 (159620) =    55bp at 81.8% exon
```

VISTA plot



VISTA plot



http://www-gsd.lbl.gov/vista

> 27000 queries on-line, distributed > 1100 copies of the program in 47 countries.

After VISTA publications at the end of 2000:

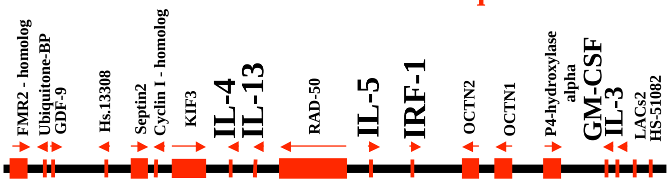
~60 papers cited VISTA and presented results obtained with the program

Biological story

Discovering Interleukin Expression Switch

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science. 2000 Apr 7;288(5463):136-40.

IL Cluster HUM 5q31



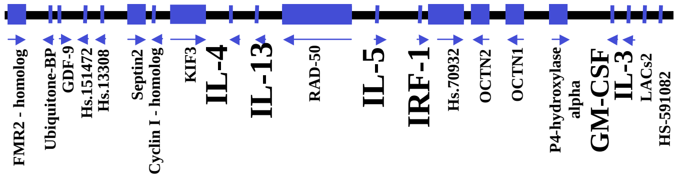
Coding

Exons 3%

Non-Coding

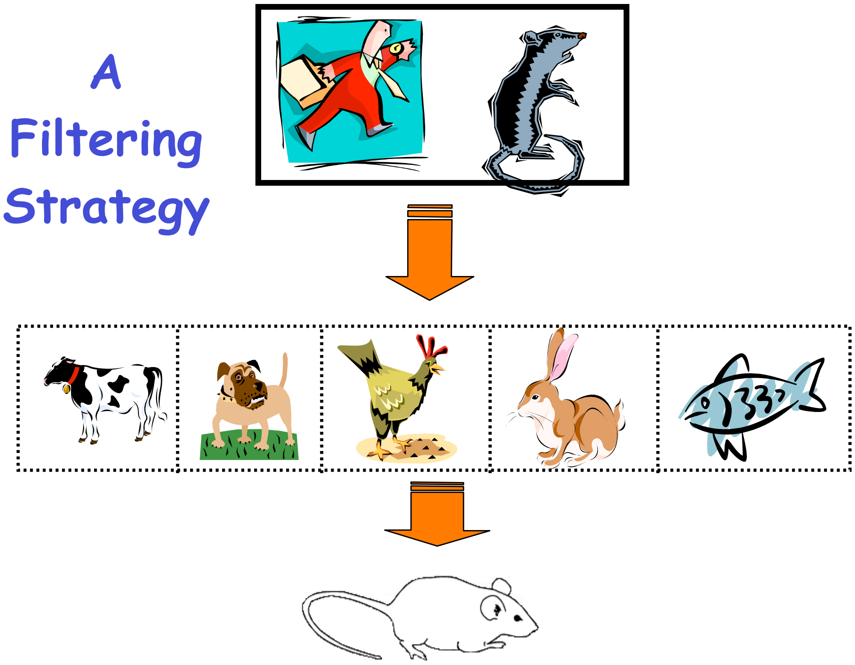
Conserved 2.6%

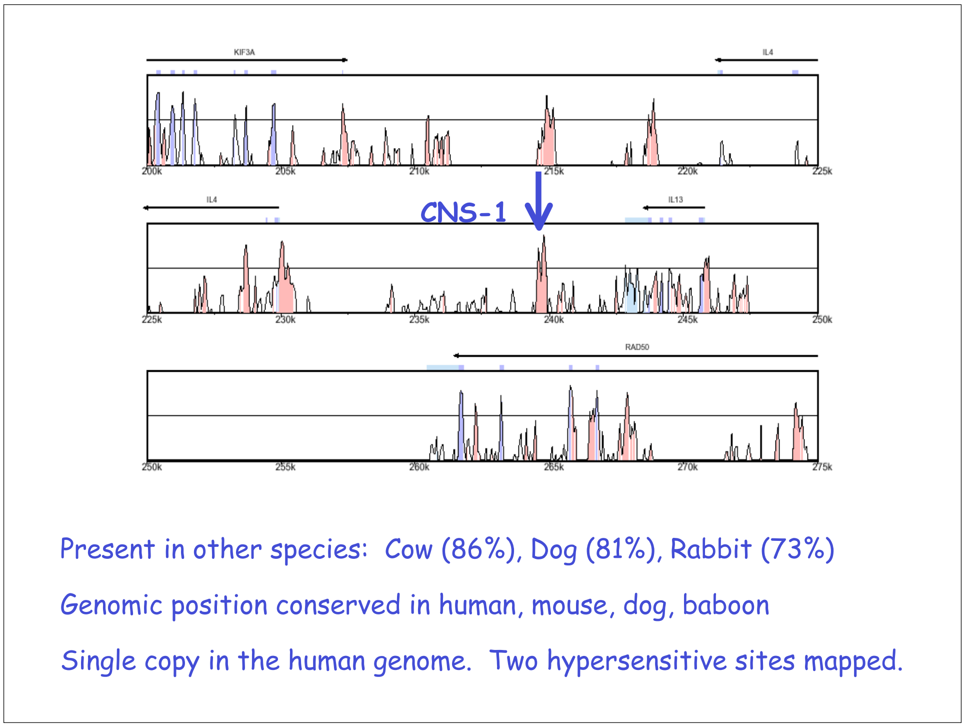
(>100bp > 75%)



IL Cluster MU Ch 11

A Filtering Strategy

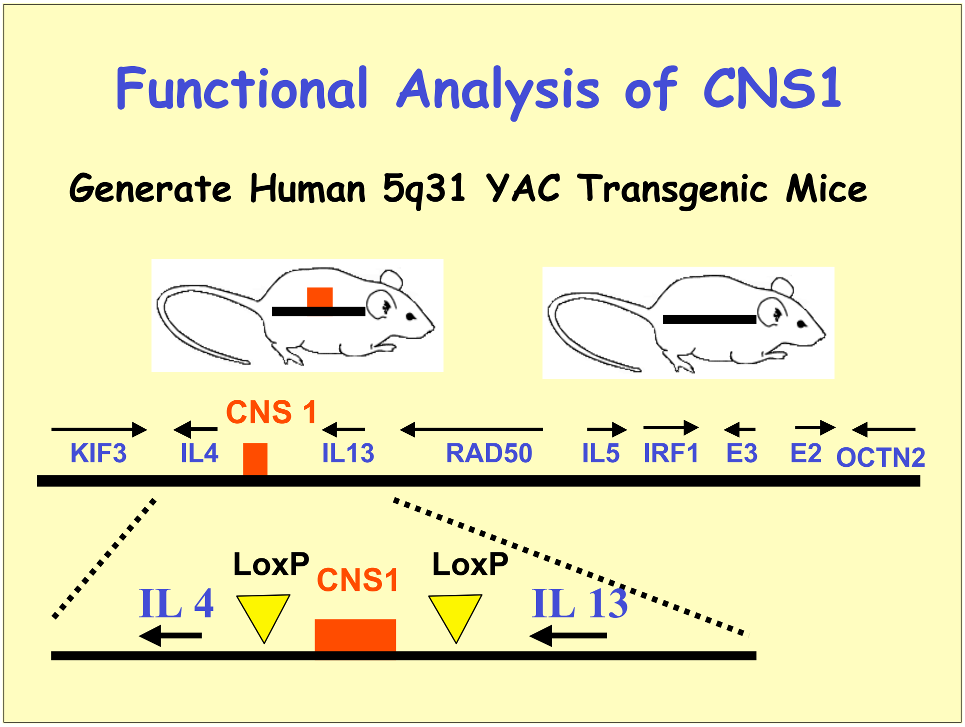




Present in other species: Cow (86%), Dog (81%), Rabbit (73%)

Genomic position conserved in human, mouse, dog, baboon

Single copy in the human genome. Two hypersensitive sites mapped.

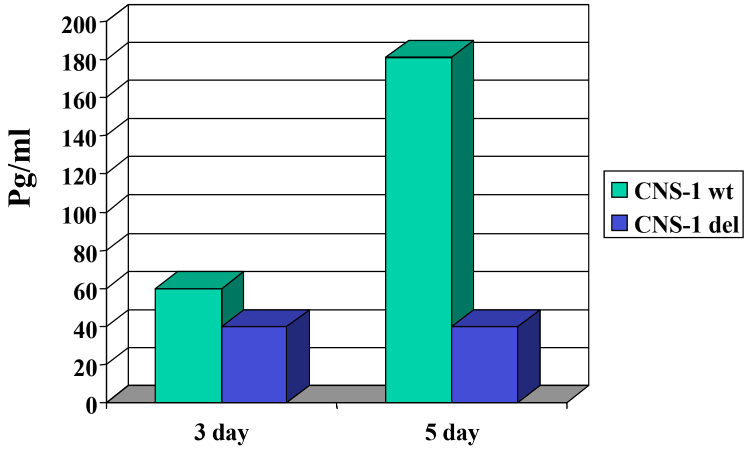


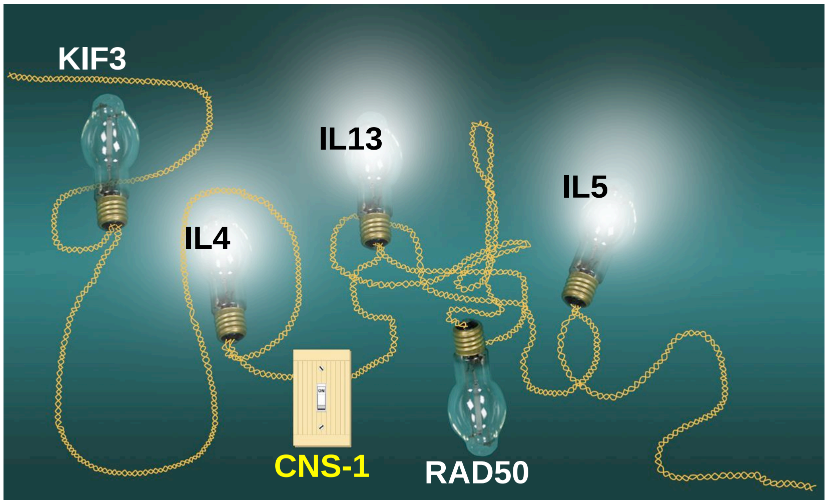
Functional Analysis of CNS1

Generate Human 5q31 YAC Transgenic Mice

Human IL 4 Production in YAC Transgenics Containing and Lacking CNS1

IL-5 & IL13 Expression is also reduced in CNS-1^{del} mice



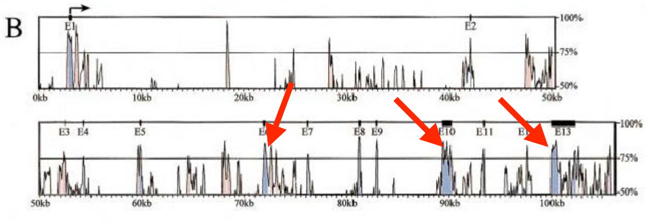


Results obtained with VISTA

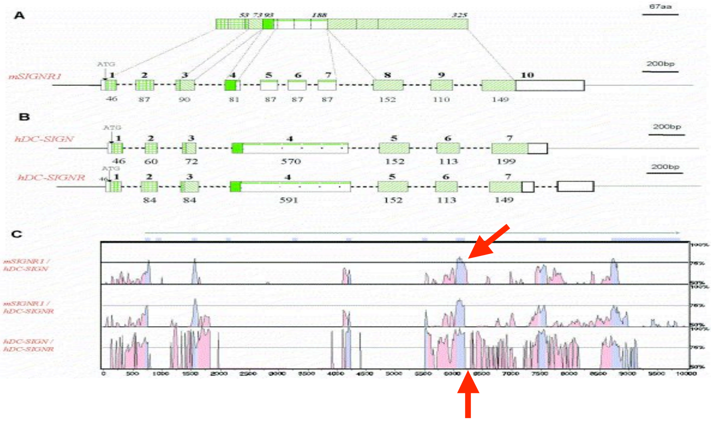
J Mol Cell Cardiol 34, 1345-1356 (2002)
Myocardin: A Component of a Molecular Switch for Smooth Muscle Differentiation. J. Chen, C. M. Kitchen, J. W. Streb and J. M. Miano

University of Oxford

VSTA used to solve the gene structures of rat and human myocardin.



Gene 293, 33–46 (2002)
Molecular characterization of the murine SIGNR1 gene encoding a C-type lectin homologous to human DC-SIGN and DC-SIGNR
S. A. Parent, T. Zhang, G. Chrebet, J. A. Clemas, D. J. Figueroa, B. Ky, R. A. Blevins, C. P. Austin and H. Rosen
Merc

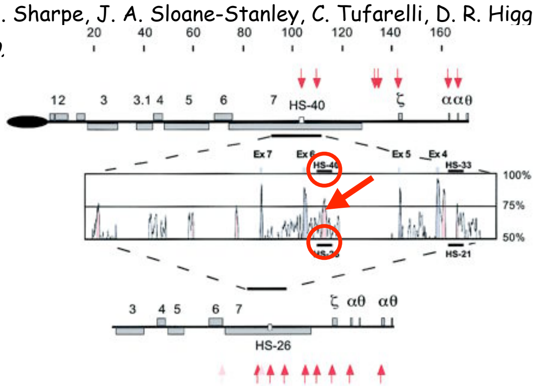


Blood, 100, 3450-3456 (2002)

Deletion of the mouse a -globin regulatory element (HS 26) has an unexpectedly mild phenotype

E. Anguita, J. A. Sharpe, J. A. Sloane-Stanley, C. Tufarelli, D. R. Higgs, and W. G. Wood

University of O.



(HS 40) is necessary for high-level expression of the a-globin genes. A similar element in the mouse (mHS 26) supposedly has similar functional properties. Knock out mHS26 instead of the expected severe a -thalassemia phenotype, produce the mice with a mild disease. These results may indicate differences in the regulation of the a -globin clusters in mice and humans.

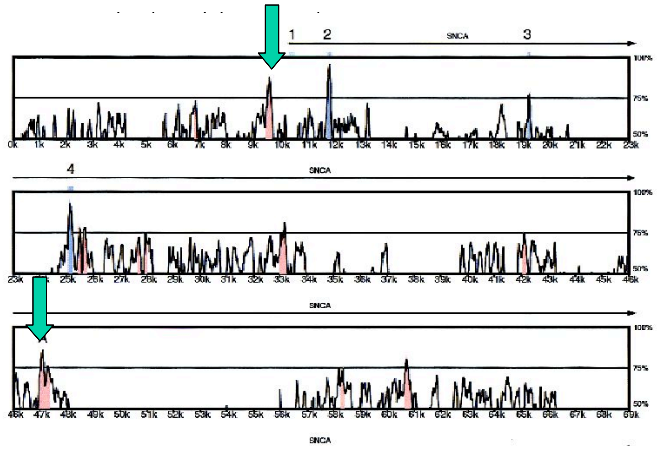
Genome Research 11, 78 (2001)

Human and Mouse - Synuclein Genes: Comparative Genomic Sequence Analysis and Identification of a Novel Gene Regulatory Element

J. W. Touchman, et al.

NIH Intramural Sequencing Center, National Institutes of Health

Syı



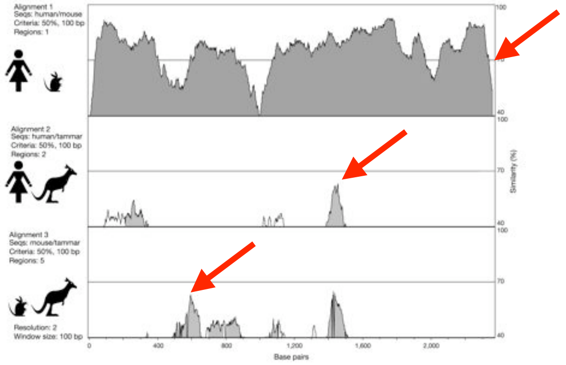
EMBO reports 4:143 (2003)

The kangaroo genome. Leaps and bounds in comparative genomics

M. J. Wakefield and J. A. Marshall Graves

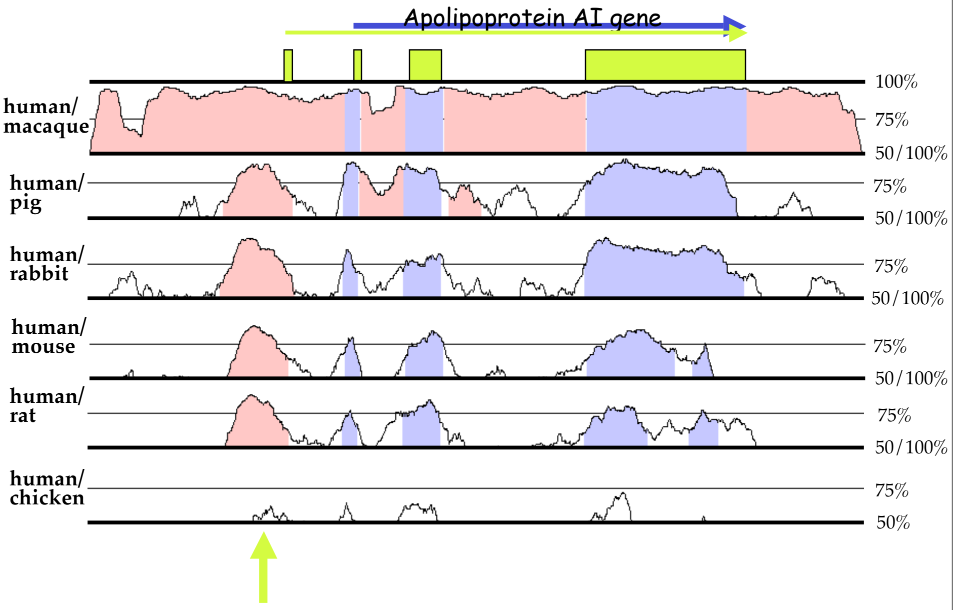
Research School of Biological Sciences, The Australian National University,
Canberra, ACT 0200, Australia

'The kangaroo genome is a rich and unique resource for comparative genomics, a treasure trove of comparative genomics data'.



Phylogenetic footprinting of 3′ untranslated region of the SLC16A2 gene

Multi-Species Comparative Analysis (VISTA)



Apolipoprotein AI gene

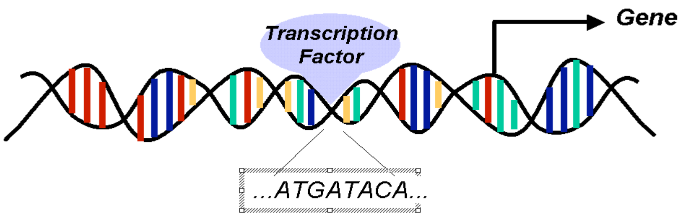
Liver enhancer

VISTA family of tools

http://www-gsd.lbl.gov/vista

- VISTA – comparing DNA of multiple organisms
- for 3 species - analyzing cutoffs to define actively conserved non-coding sequences
- cVISTA - comparing two closely related species
- rVISTA – regulatory VISTA

Identifying non-coding sequences (CNSs) involved in transcriptional regulation



rVISTA - prediction of transcription factor binding sites

- Simultaneous searches of the major transcription factor binding site database (Transfac) and the use of global sequence alignment to sieve through the data

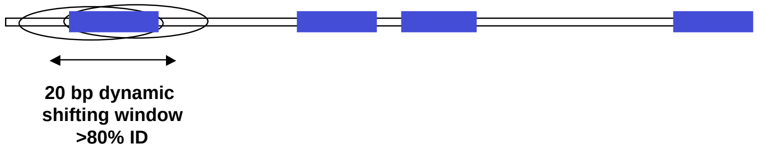
- Combination of database searches with comparative sequence analysis reduces the number of predicted transcription factor binding sites by several orders of magnitude

Regulatory VISTA (rVISTA)

1. Identify potential transcription factor binding sites for each sequence using library of matrices (TRANSFAC)
2. Identify aligned sites using VISTA
3. Identify conserved sites using dynamic shifting window

Percentage of conserved sites of the total 3-5%

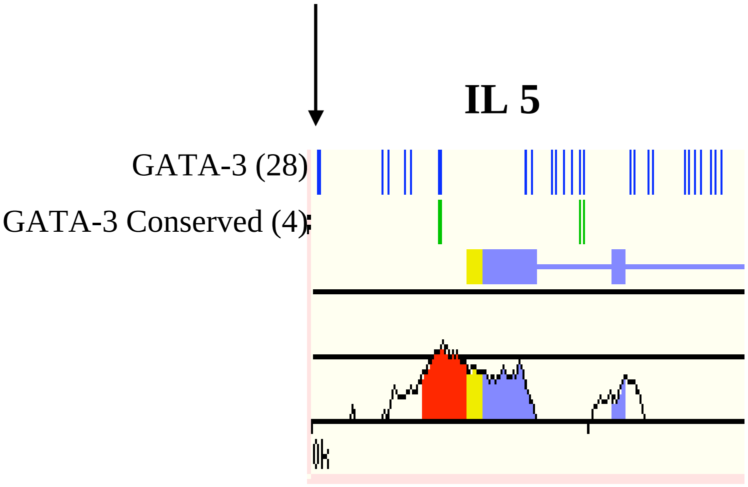
```
                 Ikaros-2                    Ikaros-2   NFAT                     Ikaros-2
Human  TGATTTCTCGGCAGCAAGGGAGGGCCCCATGACAAAGCCATTTGAAATCCCAGAAGCAATTTTCTACTTACGACCTCACTTTCTGTTGCTGTCTCTCCCTTCC
Mouse  TGATTTCTCGGCAGCCAGGGAGGGCCCCATGACGAAGCCACTCGAAATCCCAGAAGCAATTTTCTACTTACGACCTCACTTTCTGTTGCTCTCTCTTCCTCCC
Dog    TGATTTCTCGGCAGCAAGGGAGGGCCCCATGACGAAGCCATTTGAAATCCCAGAAGCGATTTTCTACCTACGACCTCACTTTCTGTTGCGCTCACTCCCTTCCC
Rat    TGATTTCTCGGCAGCCAGGGAGGGCCCCATGACGAAGCCACTCGAAATCCCAGAAGCAATTTTCTACTTACGACCTCACTTTCTGTTGTTCTCTCTTCCTCCCC
Cow    TGATTTCTCGGCAGCCAGGGAGGGCCCCATGACGAAGCCATTTGAAATCCCAGAAGCAATTTTCTACTTACGACCTCACTTTCTGTTGCGTTCTCTCCCTTCCC
Rabbit TGATTTCTCGGCAGCCAGGGAGGGCCCCACGAC-AAGCCATTCAAAATCCCAGAAGTGATTTTCTACTTACGACCTCACTTTCTGTTG----CTCTCTCCTTCCCT
```



~1 Meg region, 5q31

	Coding	Noncoding
▯		
Human interval Transfac predictions for GATA sites	839	20654
▯		
Aligned with the same predicted site in the mouse seq.	450	2618
Alligned sites conserved at 80% / 24 bp dynamic window	303	731
▯		
Random DNA sequence of the same length		29280▯

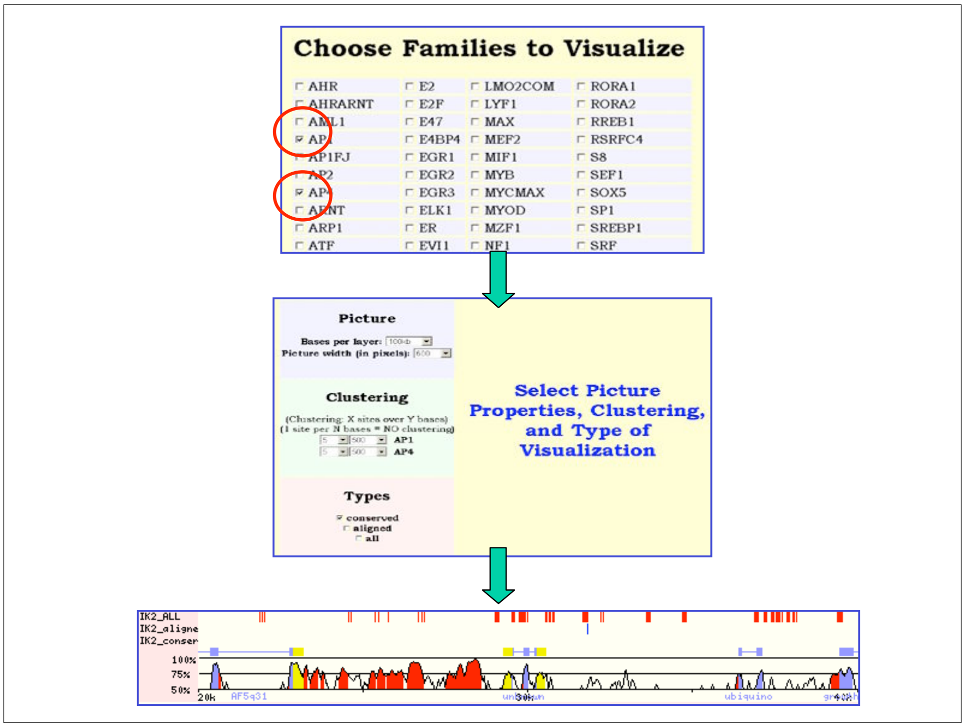
2 Exp. Verified GATA-3 Sites



IL 5

GATA-3 (28)

GATA-3 Conserved (4)



Choose Families to Visualize

- □ AHR
- □ AHRARNT
- □ AML1
- ☑ AP1
- □ AP1FJ
- □ AP2
- ☑ AP4
- □ ARNT
- □ ARP1
- □ ATF
- □ E2
- □ E2F
- □ E47
- □ E4BP4
- □ EGR1
- □ EGR2
- □ EGR3
- □ ELK1
- □ ER
- □ EVI1
- □ LMO2COM
- □ LYF1
- □ MAX
- □ MEF2
- □ MIF1
- □ MYB
- □ MYCMAX
- □ MYOD
- □ MZF1
- □ NF1
- □ RORA1
- □ RORA2
- □ RREB1
- □ RSRFC4
- □ S8
- □ SEF1
- □ SOX5
- □ SP1
- □ SREBP1
- □ SRF

Picture

Bases per layer: 100kb

Picture width (in pixels): 600

Clustering

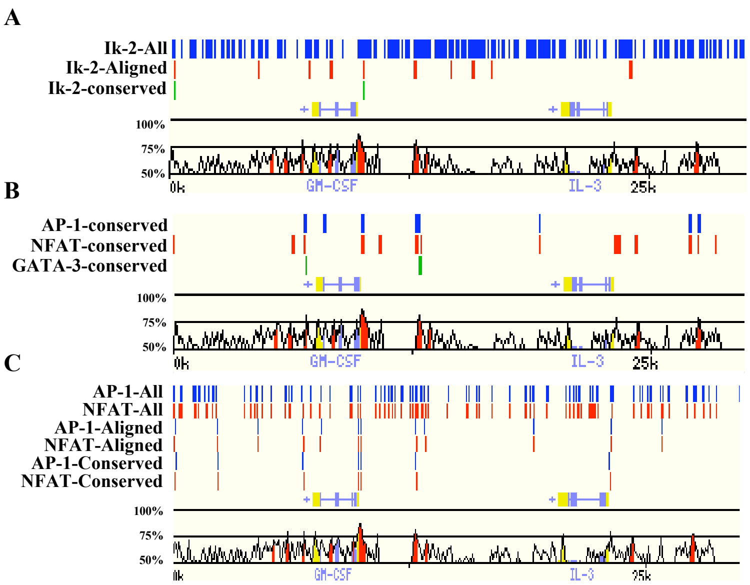
(Clustering: X sites over Y bases)
(1 site per N bases = NO clustering)

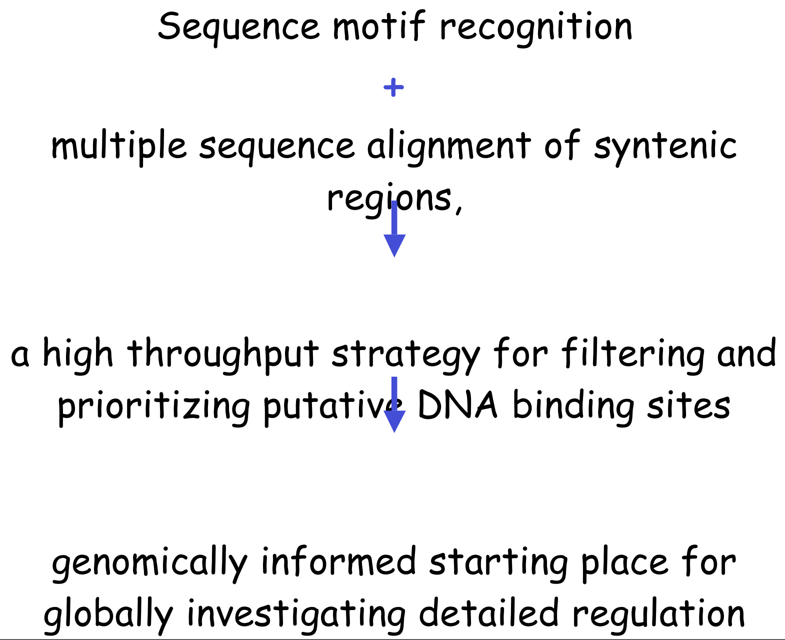
5 500 AP1
5 500 AP4

Types

☑ conserved
□ aligned
□ all

Select Picture Properties, Clustering, and Type of Visualization





Sequence motif recognition

+

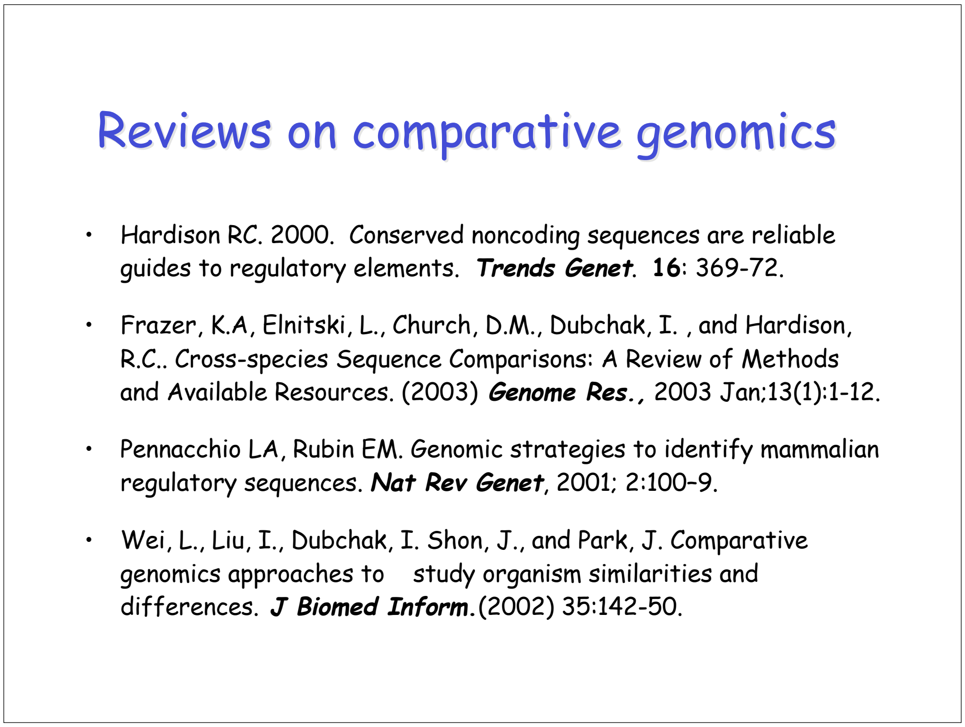
multiple sequence alignment of syntenic regions,

a high throughput strategy for filtering and prioritizing putative DNA binding sites

genomically informed starting place for globally investigating detailed regulation

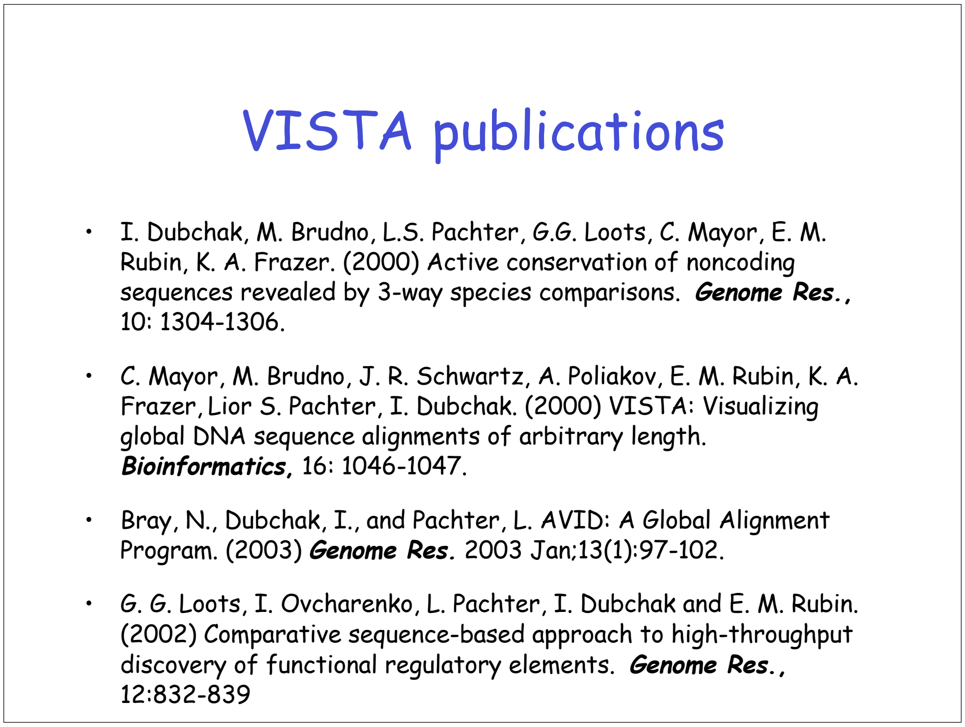
Main features of VISTA

- Clear , configurable output
- Ability to visualize several global alignments on the same scale
- Alignments up to several megabases
- Working with finished and draft sequences
- Available source code and WEB site



Reviews on comparative genomics

- Hardison RC. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. ***Trends Genet***. **16**: 369-72.
- Frazer, K.A, Elnitski, L., Church, D.M., Dubchak, I. , and Hardison, R.C.. Cross-species Sequence Comparisons: A Review of Methods and Available Resources. (2003) ***Genome Res.,*** 2003 Jan;13(1):1-12.
- Pennacchio LA, Rubin EM. Genomic strategies to identify mammalian regulatory sequences. ***Nat Rev Genet***, 2001; 2:100–9.
- Wei, L., Liu, I., Dubchak, I. Shon, J., and Park, J. Comparative genomics approaches to study organism similarities and differences. ***J Biomed Inform.***(2002) 35:142-50.



VISTA publications

- I. Dubchak, M. Brudno, L.S. Pachter, G.G. Loots, C. Mayor, E. M. Rubin, K. A. Frazer. (2000) Active conservation of noncoding sequences revealed by 3-way species comparisons. ***Genome Res.,*** 10: 1304-1306.
- C. Mayor, M. Brudno, J. R. Schwartz, A. Poliakov, E. M. Rubin, K. A. Frazer, Lior S. Pachter, I. Dubchak. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length. ***Bioinformatics,*** 16: 1046-1047.
- Bray, N., Dubchak, I., and Pachter, L. AVID: A Global Alignment Program. (2003) ***Genome Res.*** 2003 Jan;13(1):97-102.
- G. G. Loots, I. Ovcharenko, L. Pachter, I. Dubchak and E. M. Rubin. (2002) Comparative sequence-based approach to high-throughput discovery of functional regulatory elements. ***Genome Res.,*** 12:832-839

What if you don't have sequences of different species for the genomic region of your interest?

Are there publicly available comparative genomics data?

Large scale VISTA applications:

The Berkeley Genome Pipeline – comparing complete genomes
http://pipeline.lbl.gov

Cardiovascular comparative genomics database
http://pga.lbl.gov

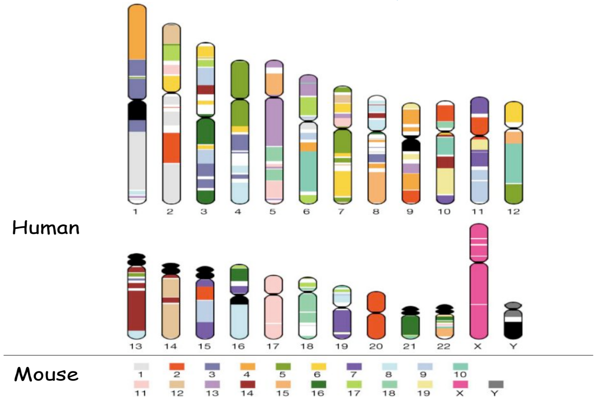
Development of automatic computational system for comparative analysis of whole genomes

2001 – Whole mouse genome assemblies became available
Human genome – high quality draft

Precomputed alignments:
Human Genome (Golden Path Assembly)
against
Mouse assemblies: Arachne, Phusion (2001) MGSC v3 (2002)
Rat assemblies: January 2003, February 2003

D.Melanogaster vs D.Pseudoobscura February 2003

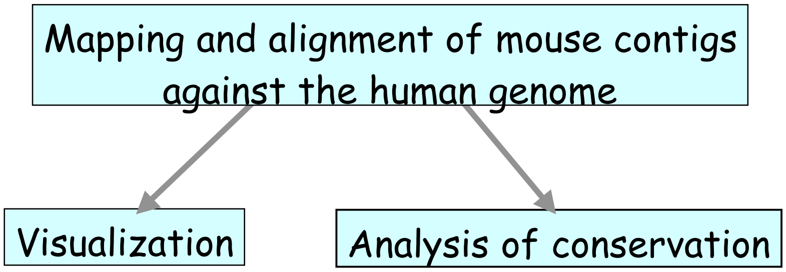
Chromosome Comparison



Base pair alignment

```
247 GGTGAGGTCGAGGACCCTGCA  CGGAGCTGTATGGAGGGCA   AGAGC
    |:   ||  ||||:  |||| --:||  ||| |::|   |||---||||
368 GAGTCGGGGGAGGGGGCTGCTGTTGGCTCTGGACAGCTTGCATTGAGAGG
```

Main modules of the system



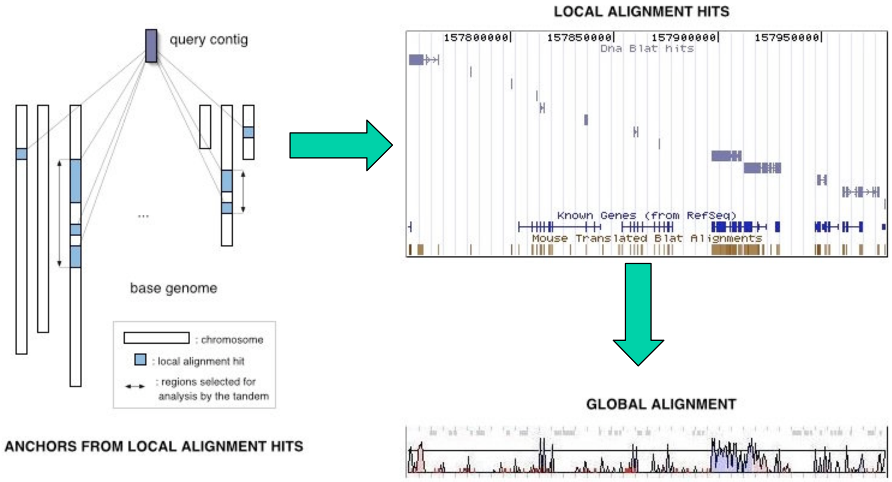
Mapping and alignment of mouse contigs against the human genome

Visualization

Analysis of conservation

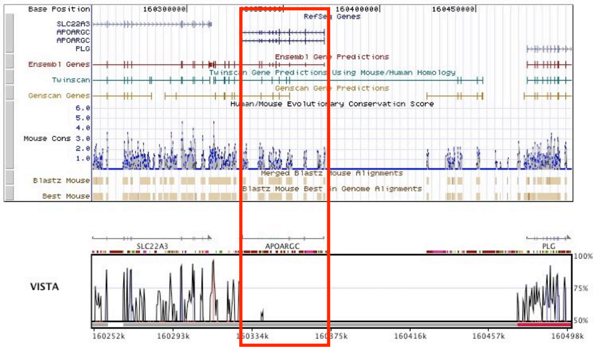
Tandem Local/Global Alignment Approach

Sequence fragment anchoring (DNA and/or translated BLAT)
Multi-step verification of potential regions using global alignment (AVID or LAGAN)



Tandem approach in comparison with local alignment

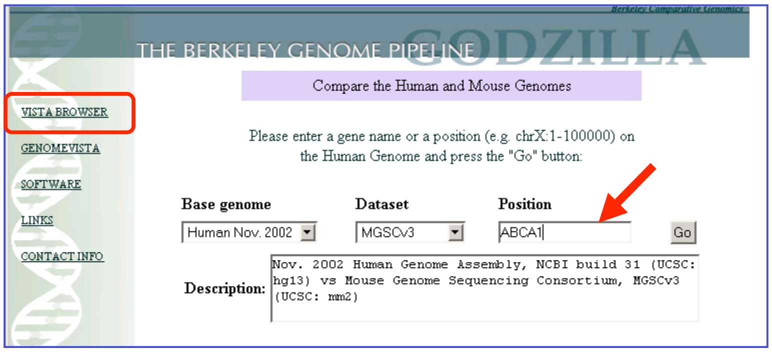
Better specificity while preserving good sensitivity



Apolipoprotein(a) region. The expressed gene is confined to a subset of primates. Our method predicts that apoa(a) has no homology in the mouse that local alignment can't detect.

VISTA Browser

Preprocessed whole genome comparison for pairs of species (human/mouse/rat & drosophilas)



Compare the Human and Mouse Genomes

Please enter a gene name or a position (e.g. chrX:1-100000) on the Human Genome and press the "Go" button:

Base genome: Human Nov. 2002

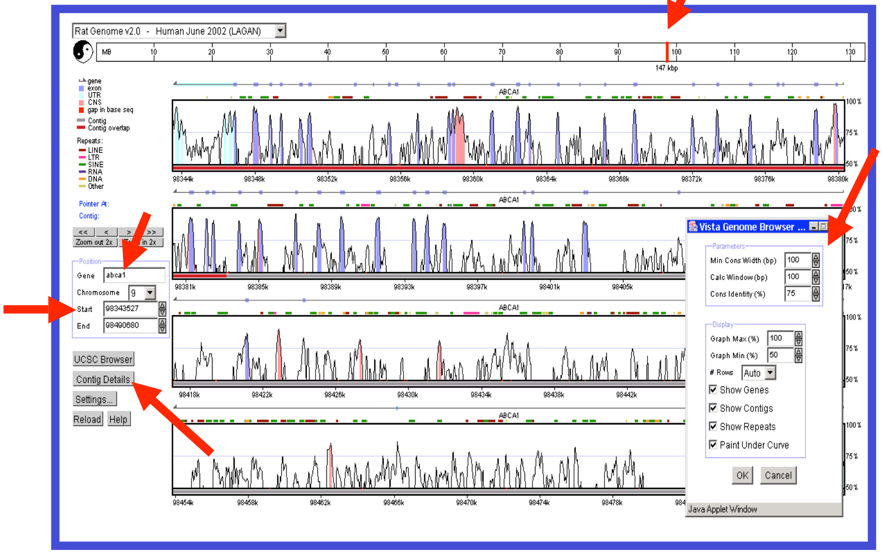
Dataset: MGSCv3

Position: ABCA1

Description: Nov. 2002 Human Genome Assembly, NCBI build 31 (UCSC: hg13) vs Mouse Genome Sequencing Consortium, MGSCv3 (UCSC: mm2)

http://pipeline.lbl.gov/

VistaBrowser



Rat Genome v2.0 - Human June 2002 (LAGAN)

Position

Gene: abca1

Chromosome: 9

Start: 98343527

End: 98490680

Parameters

Min Cons Width (bp): 100

Calc Window (bp): 100

Cons Identity (%): 75

Display

Graph Max (%): 100

Graph Min (%): 50

Rows: Auto

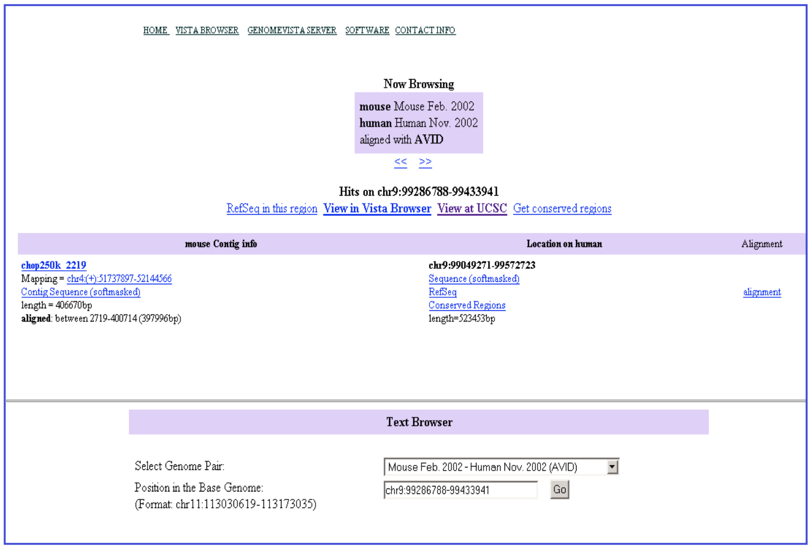
☑ Show Genes

☑ Show Contigs

☑ Show Repeats

☑ Paint Under Curve

Text browser



HOME VISTA BROWSER GENOMEVISTA SERVER SOFTWARE CONTACT INFO

Now Browsing

mouse Mouse Feb. 2002
human Human Nov. 2002
aligned with **AVID**

<< >>

Hits on chr9:99286788-99433941

RefSeq in this region **View in Vista Browser** View at UCSC Get conserved regions

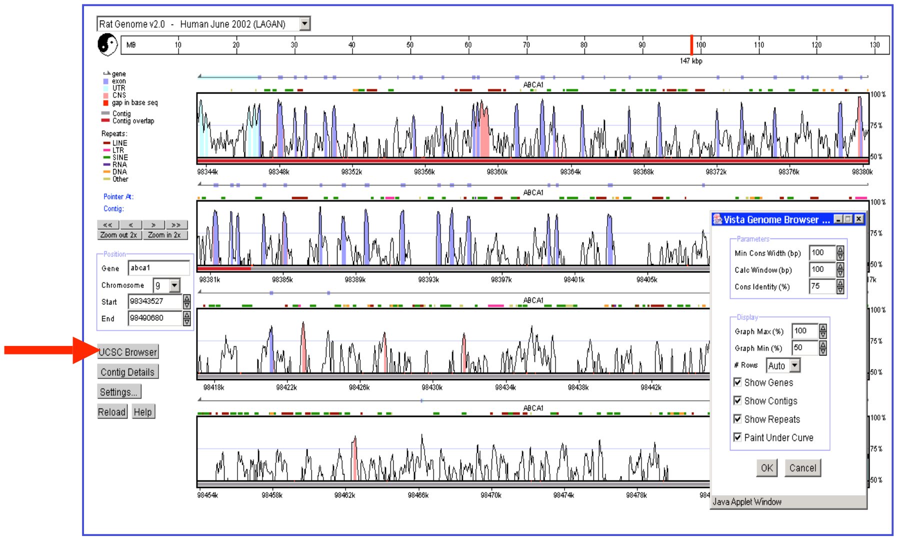
mouse Contig info	Location on human	Alignment
chop250k_2219 Mapping = chr4:(+):51737897-52144566 Contig Sequence (softmasked) length = 406670bp **aligned**: between 2719-400714 (397996bp)	**chr9:99049271-99572723** Sequence (softmasked) RefSeq Conserved Regions length=523453bp	alignment

Text Browser

Select Genome Pair: Mouse Feb. 2002 - Human Nov. 2002 (AVID)

Position in the Base Genome: chr9:99286788-99433941 Go
(Format: chr11:113030619-113173035)

VistaBrowser

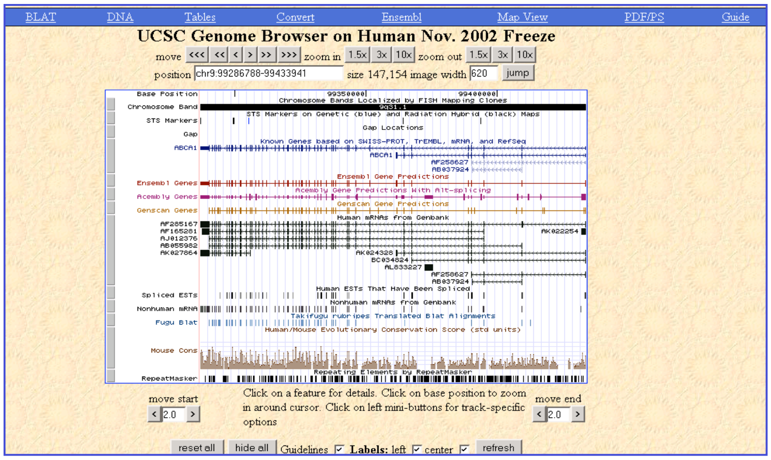


Position
Gene: abca1
Chromosome: 9
Start: 98343527
End: 98490680

Parameters
Min Cons Width (bp): 100
Calc Window (bp): 100
Cons Identity (%): 75

Display
Graph Max (%): 100
Graph Min (%): 50
Rows: Auto
☑ Show Genes
☑ Show Contigs
☑ Show Repeats
☑ Paint Under Curve

ABCA1 interval in UCSC human genome browser



UCSC Genome Browser on Human Nov. 2002 Freeze

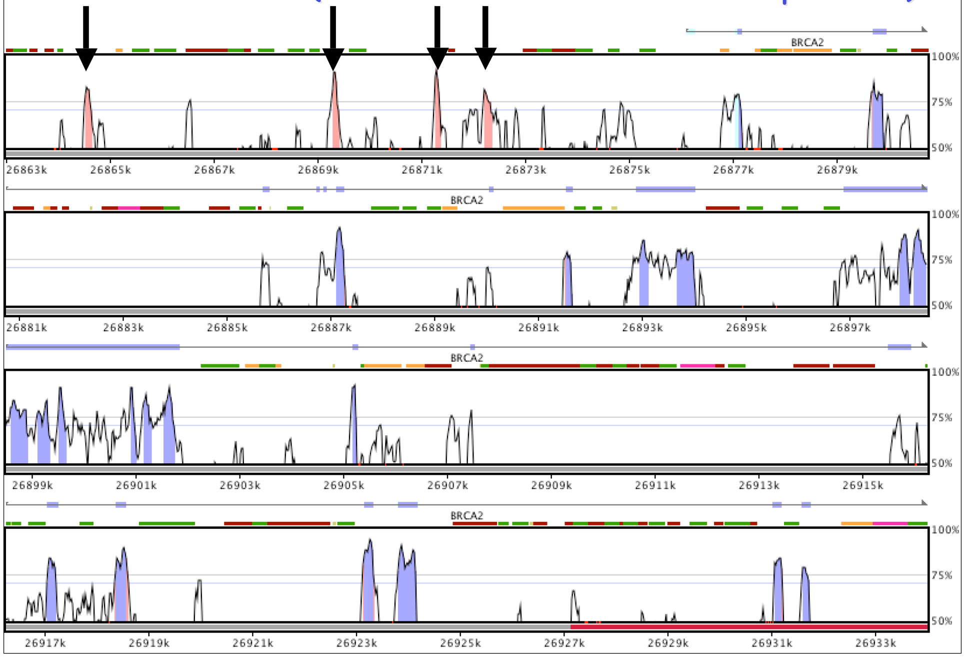
position chr9:99286788-99433941 size 147,154 image width 620

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options

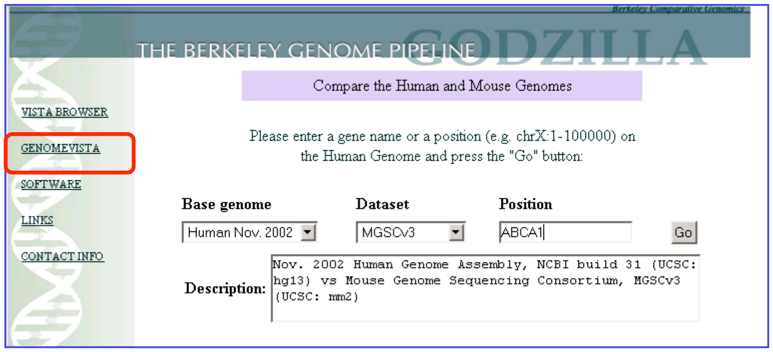
move start 2.0 move end 2.0

Guidelines ☑ **Labels:** left ☑ center ☑

VISTA Browser (Human/Mouse BRCA2 Comparison)



GenomeVista - is an interactive for comparing your favorite sequence against the base genome



Compare the Human and Mouse Genomes

Please enter a gene name or a position (e.g. chrX:1-100000) on the Human Genome and press the "Go" button:

Base genome: Human Nov. 2002

Dataset: MGSCv3

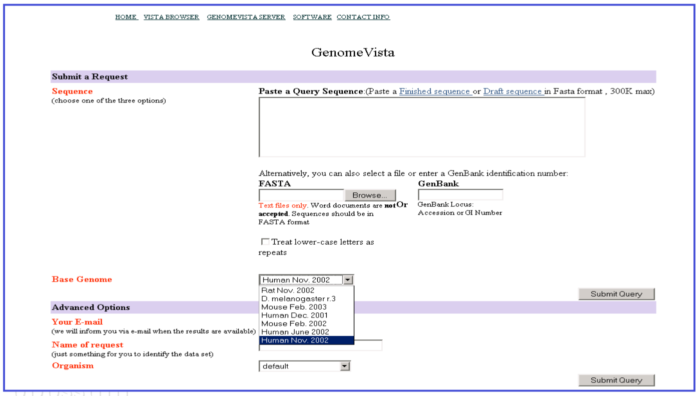
Position: ABCA1

Description: Nov. 2002 Human Genome Assembly, NCBI build 31 (UCSC: hg13) vs Mouse Genome Sequencing Consortium, MGSCv3 (UCSC: mm2)

http://pipeline.lbl.gov/

GenomeVISTA

Self-Input Sequence Comparison to either Human, Mouse, Rat, D.Melanogaster Reference Genomes



GenomeVista

Submit a Request

Sequence
(choose one of the three options)

Paste a Query Sequence:(Paste a Finished sequence or Draft sequence in Fasta format , 300K max)

Alternatively, you can also select a file or enter a GenBank identification number:

FASTA
Text files only. Word documents are **not** **accepted**. Sequences should be in FASTA format

Or

GenBank
GenBank Locus:
Accession or GI Number

□ Treat lower-case letters as repeats

Base Genome: Human Nov. 2002

Advanced Options

Your E-mail
(we will inform you via e-mail when the results are available)

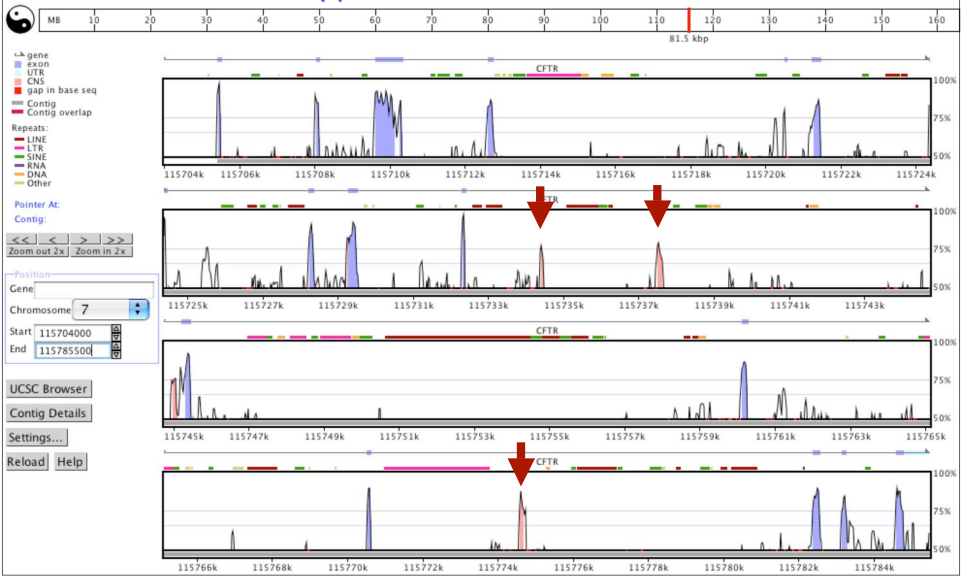
Name of request
(just something for you to identify the data set)

Organism: default

http://pipeline.lbl.gov/

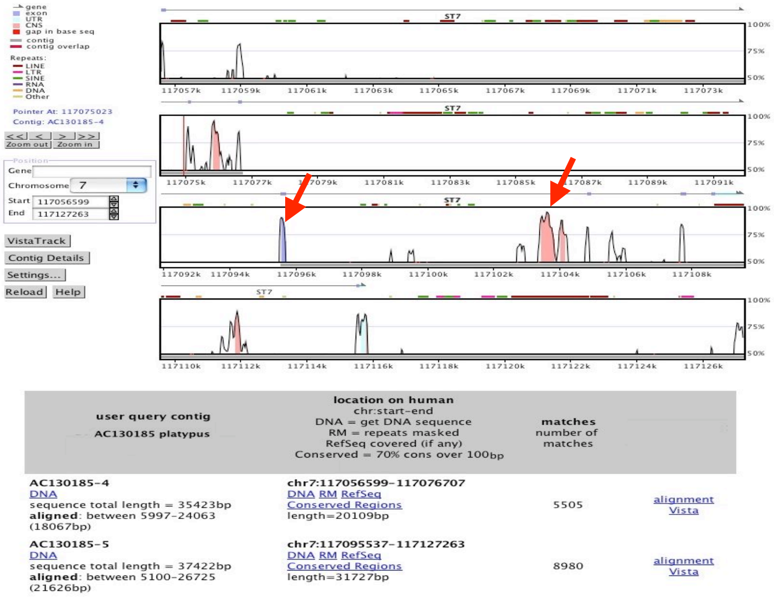
GenomeVISTA

Random Opposum BAC versus Human Genome



MB 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
81.5 kbp
gene
exon
UTR
CNS
gap in base seq
Contig
Contig overlap
Repeats:
LINE
LTR
SINE
RNA
DNA
Other
Pointer At:
Contig:
<< < > >>
Zoom out 2x
Zoom in 2x
Position
Gene
Chromosome 7
Start 115704000
End 115785500
UCSC Browser
Contig Details
Settings...
Reload
Help
CFTR
100%
75%
50%
115704k 115706k 115708k 115710k 115712k 115714k 115716k 115718k 115720k 115722k 115724k
115725k 115727k 115729k 115731k 115733k 115735k 115737k 115739k 115741k 115743k
115745k 115747k 115749k 115751k 115753k 115755k 115757k 115759k 115761k 115763k 115765k
115766k 115768k 115770k 115772k 115774k 115776k 115778k 115780k 115782k 115784k

Results of an on-line submission of a draft unannotated platypus sequence AC130185 to Genome Vista. The gene has been correctly identified.



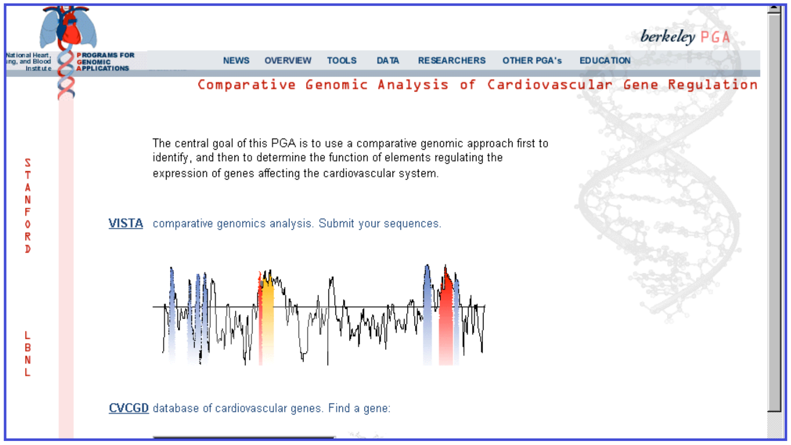
gene
exon
UTR
CNS
gap in base seq
contig
contig overlap
Repeats:
LINE
LTR
SINE
RNA
DNA
Other
Pointer At: 117075023
Contig: AC130185-4
<< < > >>
Zoom out
Zoom in
Position
Gene
Chromosome 7
Start 117056599
End 117127263
VistaTrack
Contig Details
Settings...
Reload
Help
ST7
100%
75%
50%
117057k 117059k 117061k 117063k 117065k 117067k 117069k 117071k 117073k
117075k 117077k 117079k 117081k 117083k 117085k 117087k 117089k 117091k
117092k 117094k 117096k 117098k 117100k 117102k 117104k 117106k 117108k
117110k 117112k 117114k 117116k 117118k 117120k 117122k 117124k 117126k

user query contig
AC130185 platypus
	location on human
	chr:start-end
	DNA = get DNA sequence
	RM = repeats masked
	RefSeq covered (if any)
	Conserved = 70% cons over 100bp
		matches
		number of matches	

AC130185-4 DNA sequence total length = 35423bp **aligned**: between 5997-24063 (18067bp)	**chr7:117056599-117076707** DNA RM RefSeq Conserved Regions length=20109bp	5505	alignment Vista
AC130185-5 DNA sequence total length = 37422bp **aligned**: between 5100-26725 (21626bp)	**chr7:117095537-117127263** DNA RM RefSeq Conserved Regions length=31727bp	8980	alignment Vista

Comparative analysis of genomic intervals containing important cardiovascular genes

http://pga.lbl.gov



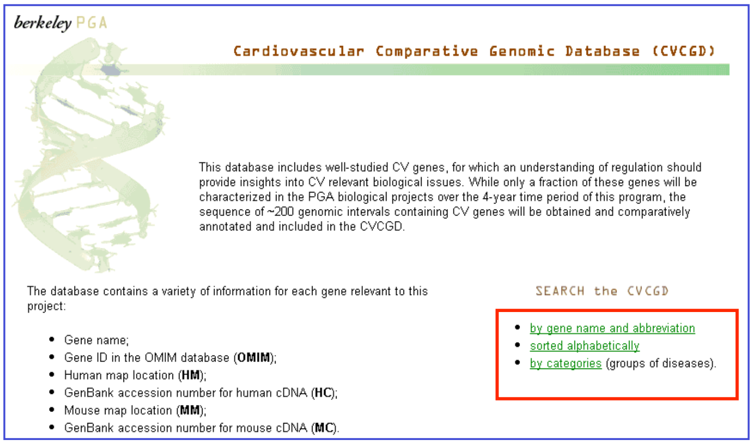
Comparative Genomic Analysis of Cardiovascular Gene Regulation

The central goal of this PGA is to use a comparative genomic approach first to identify, and then to determine the function of elements regulating the expression of genes affecting the cardiovascular system.

VISTA comparative genomics analysis. Submit your sequences.

CVCGD database of cardiovascular genes. Find a gene:

http://pga.lbl.gov/cvcgd.html



Cardiovascular Comparative Genomic Database (CVCGD)

This database includes well-studied CV genes, for which an understanding of regulation should provide insights into CV relevant biological issues. While only a fraction of these genes will be characterized in the PGA biological projects over the 4-year time period of this program, the sequence of ~200 genomic intervals containing CV genes will be obtained and comparatively annotated and included in the CVCGD.

The database contains a variety of information for each gene relevant to this project:

- Gene name;
- Gene ID in the OMIM database (**OMIM**);
- Human map location (**HM**);
- GenBank accession number for human cDNA (**HC**);
- Mouse map location (**MM**);
- GenBank accession number for mouse cDNA (**MC**).

SEARCH the CVCGD

- by gene name and abbreviation
- sorted alphabetically
- by categories (groups of diseases).

Search Results

Links to whole genome alignment

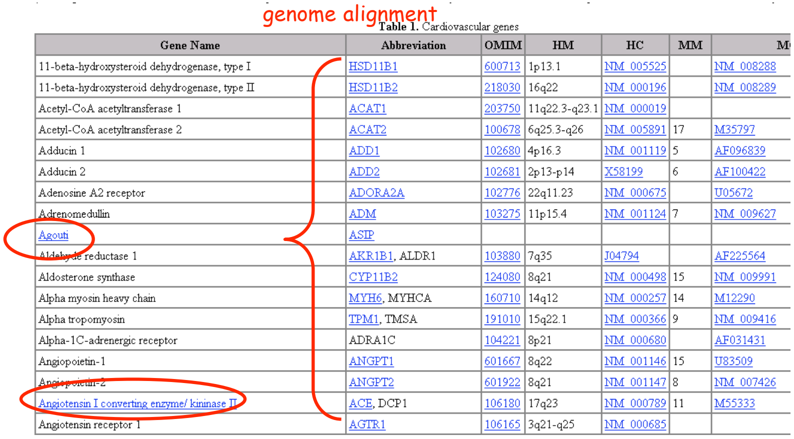
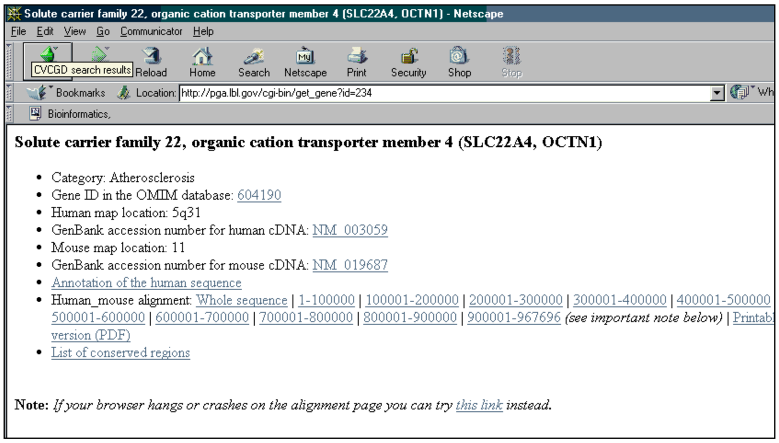


Table 1. Cardiovascular genes

Gene Name	Abbreviation	OMIM	HM	HC	MM	M
11-beta-hydroxysteroid dehydrogenase, type I	HSD11B1	600713	1p13.1	NM_005525		NM_008288
11-beta-hydroxysteroid dehydrogenase, type II	HSD11B2	218030	16q22	NM_000196		NM_008289
Acetyl-CoA acetyltransferase 1	ACAT1	203750	11q22.3-q23.1	NM_000019		
Acetyl-CoA acetyltransferase 2	ACAT2	100678	6q25.3-q26	NM_005891	17	M35797
Adducin 1	ADD1	102680	4p16.3	NM_001119	5	AF096839
Adducin 2	ADD2	102681	2p13-p14	X58199	6	AF100422
Adenosine A2 receptor	ADORA2A	102776	22q11.23	NM_000675		U05672
Adrenomedullin	ADM	103275	11p15.4	NM_001124	7	NM_009627
Agouti	ASIP					
Aldehyde reductase 1	AKR1B1, ALDR1	103880	7q35	J04794		AF225564
Aldosterone synthase	CYP11B2	124080	8q21	NM_000498	15	NM_009991
Alpha myosin heavy chain	MYH6, MYHCA	160710	14q12	NM_000257	14	M12290
Alpha tropomyosin	TPM1, TMSA	191010	15q22.1	NM_000366	9	NM_009416
Alpha-1C-adrenergic receptor	ADRA1C	104221	8p21	NM_000680		AF031431
Angiopoietin-1	ANGPT1	601667	8q22	NM_001146	15	U83509
Angiopoietin-2	ANGPT2	601922	8q21	NM_001147	8	NM_007426
Angiotensin I converting enzyme/ kininase II	ACE, DCP1	106180	17q23	NM_000789	11	M55333
Angiotensin receptor 1	AGTR1	106165	3q21-q25	NM_000685		

Sequenced in Berkeley PGA

Example of CVCGD interval sequenced in Berkeley PGA



Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1) - Netscape
File Edit View Go Communicator Help
CVCGD search results
Reload
Home
Search
Netscape
Print
Security
Shop
Stop
Bookmarks
Location:
Wh
Bioinformatics,

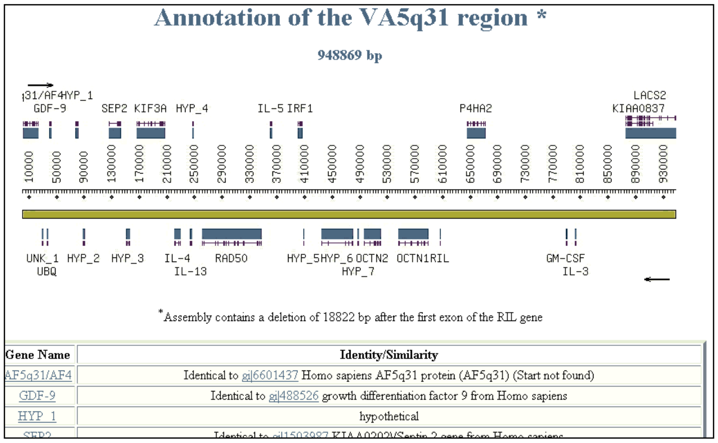
http://pga.lbl.gov/cgi-bin/get_gene?id=234

Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1)

- Category: Atherosclerosis
- Gene ID in the OMIM database: 604190
- Human map location: 5q31
- GenBank accession number for human cDNA: NM_003059
- Mouse map location: 11
- GenBank accession number for mouse cDNA: NM_019687
- Annotation of the human sequence
- Human_mouse alignment: Whole sequence | 1-100000 | 100001-200000 | 200001-300000 | 300001-400000 | 400001-500000 500001-600000 | 600001-700000 | 700001-800000 | 800001-900000 | 900001-967696 *(see important note below)* | Printabl version (PDF)
- List of conserved regions

Note: *If your browser hangs or crashes on the alignment page you can try* this link *instead.*

Short annotation of the region



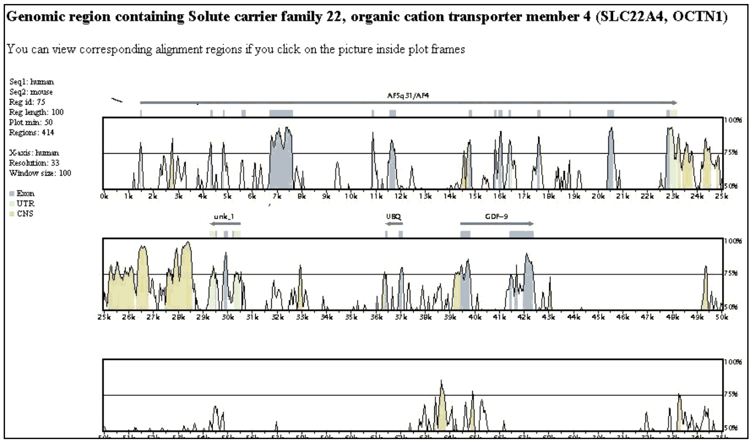
Annotation of the VA5q31 region *

948869 bp

*Assembly contains a deletion of 18822 bp after the first exon of the RIL gene

Gene Name	Identity/Similarity
AF5q31/AF4	Identical to gi|6601437 Homo sapiens AF5q31 protein (AF5q31) (Start not found)
GDF-9	Identical to gi|488526 growth differentiation factor 9 from Homo sapiens
HYP_1	hypothetical
SEP2	Identical to gi|1503987 KIAA0202\Septin 2 gene from Homo sapiens

VISTA plot of the region



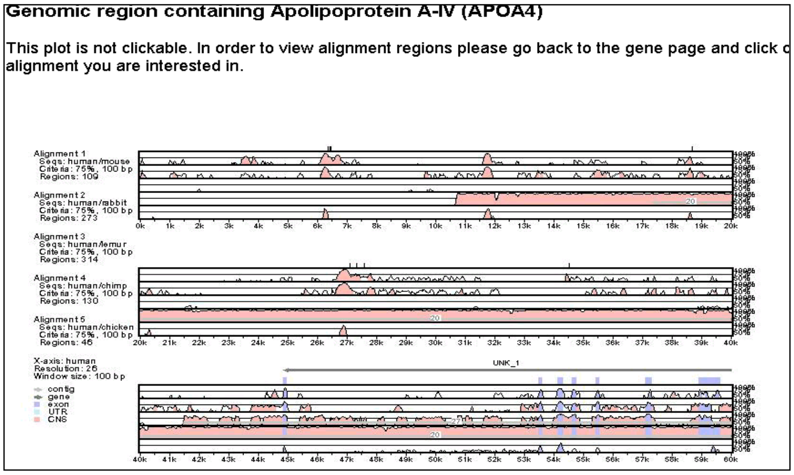
Genomic region containing Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1)

You can view corresponding alignment regions if you click on the picture inside plot frames

Seq1: human
Seq2: mouse
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 414

X-axis: human
Resolution: 33
Window size: 100

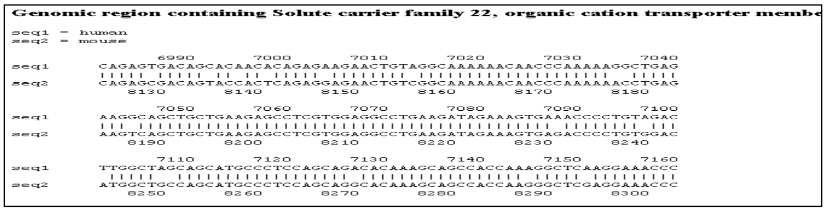
multiVISTA plot of the region



Genomic region containing Apolipoprotein A-IV (APOA4)

This plot is not clickable. In order to view alignment regions please go back to the gene page and click c alignment you are interested in.

Alignment



Genomic region containing Solute carrier family 22, organic cation transporter membe

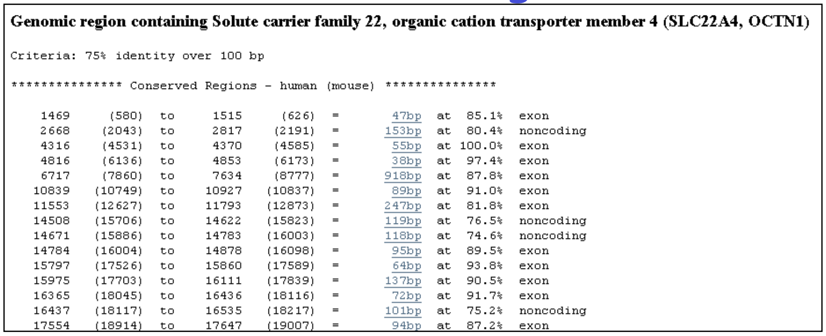
```
seq1  =  human
seq2  =  mouse

              6990      7000      7010      7020      7030      7040
seq1     CAGAGTGACAGCACAACACAGAGAAGAACTGTAGGCAAAAAACAACCCAAAAAGGCTGAG
         ||||| ||||| || || ||||| |||||||| ||||||||||||||||||||  |||||
seq2     CAGAGCGACAGTACCACTCAGAGGAGAACTGTCGGCAAAAAACAACCCAAAAAACCTGAG
            8130      8140      8150      8160      8170      8180

              7050      7060      7070      7080      7090      7100
seq1     AAGGCAGCTGCTGAAGAGCCTCGTGGAGGCCTGAAGATAGAAAGTGAAACCCCTGTAGAC
         ||| ||||||||||||||||||||||||||||||||||||||||||| |||||||| |||
seq2     AAGTCAGCTGCTGAAGAGCCTCGTGGAGGCCTGAAGATAGAAAGTGAGACCCCTGTGGAC
            8190      8200      8210      8220      8230      8240

              7110      7120      7130      7140      7150      7160
seq1     TTGGCTAGCAGCATGCCCTCCAGCAGACACAAAGCAGCCACCAAAGGCTCAAGGAAACCC
          |||||  |||||||||||||||||| ||||||||||||||||| ||||| |||||||||
seq2     ATGGCTGCCAGCATGCCCTCCAGCAGGCACAAAGCAGCCACCAAGGGCTCGAGGAAACCC
            8250      8260      8270      8280      8290      8300
```

Conserved regions

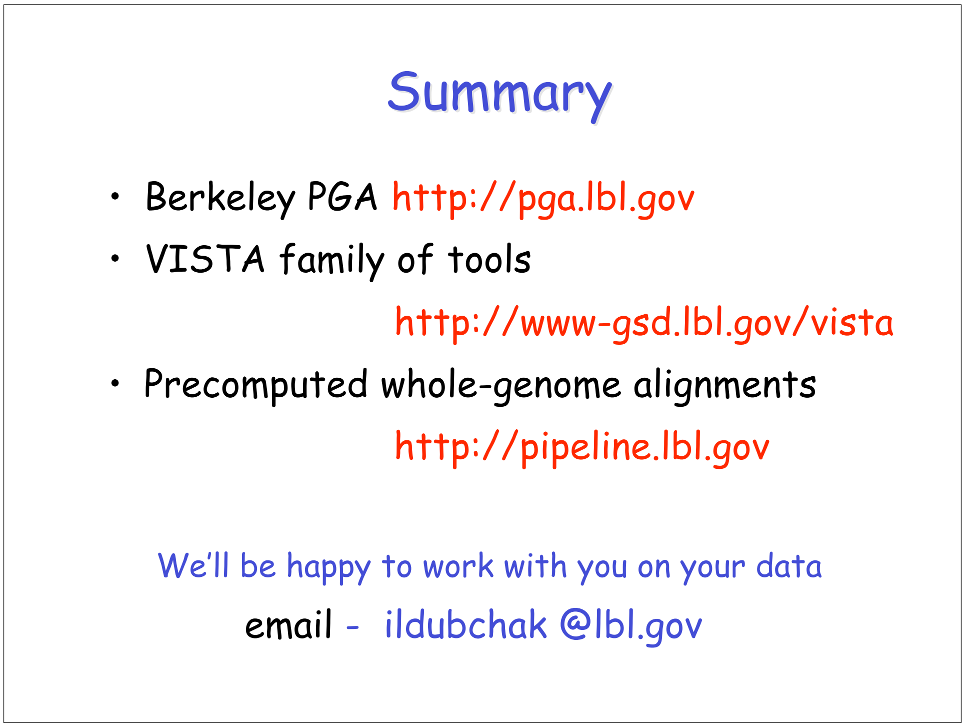


Genomic region containing Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1)

```
Criteria: 75% identity over 100 bp

*************** Conserved Regions - human (mouse) ***************

    1469     (580)  to    1515     (626)  =      47bp  at  85.1%  exon
    2668    (2043)  to    2817    (2191)  =     153bp  at  80.4%  noncoding
    4316    (4531)  to    4370    (4585)  =      55bp  at 100.0%  exon
    4816    (6136)  to    4853    (6173)  =      38bp  at  97.4%  exon
    6717    (7860)  to    7634    (8777)  =     918bp  at  87.8%  exon
   10839   (10749)  to   10927   (10837)  =      89bp  at  91.0%  exon
   11553   (12627)  to   11793   (12873)  =     247bp  at  81.8%  exon
   14508   (15706)  to   14622   (15823)  =     119bp  at  76.5%  noncoding
   14671   (15886)  to   14783   (16003)  =     118bp  at  74.6%  noncoding
   14784   (16004)  to   14878   (16098)  =      95bp  at  89.5%  exon
   15797   (17526)  to   15860   (17589)  =      64bp  at  93.8%  exon
   15975   (17703)  to   16111   (17839)  =     137bp  at  90.5%  exon
   16365   (18045)  to   16436   (18116)  =      72bp  at  91.7%  exon
   16437   (18117)  to   16535   (18217)  =     101bp  at  75.2%  noncoding
   17554   (18914)  to   17647   (19007)  =      94bp  at  87.2%  exon
```



Summary

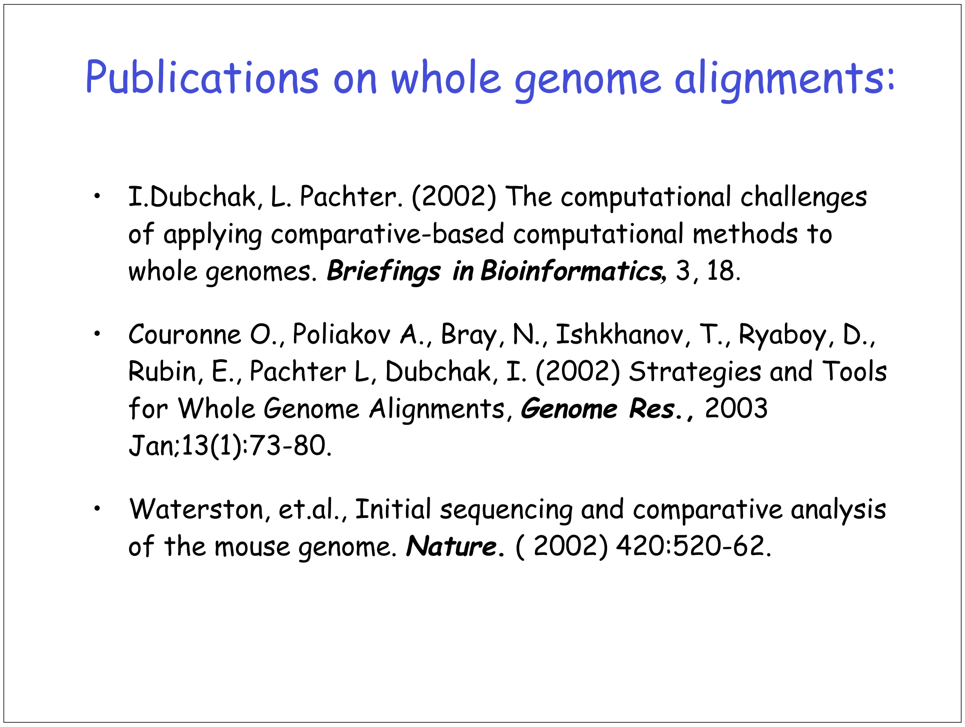
- Berkeley PGA http://pga.lbl.gov
- VISTA family of tools

  http://www-gsd.lbl.gov/vista
- Precomputed whole-genome alignments

  http://pipeline.lbl.gov

We'll be happy to work with you on your data

email - ildubchak @lbl.gov



Publications on whole genome alignments:

- I.Dubchak, L. Pachter. (2002) The computational challenges of applying comparative-based computational methods to whole genomes. ***Briefings in Bioinformatics***, 3, 18.
- Couronne O., Poliakov A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter L, Dubchak, I. (2002) Strategies and Tools for Whole Genome Alignments, ***Genome Res.***, 2003 Jan;13(1):73-80.
- Waterston, et.al., Initial sequencing and comparative analysis of the mouse genome. ***Nature.*** (2002) 420:520-62.

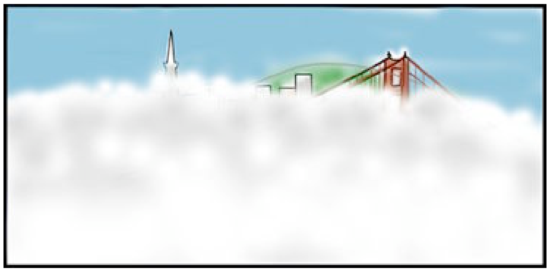
Related sites

- The Human Genome Browser & BLAT program
  http://genome.ucsc.edu/
- ENSEMBLE Project (Sanger Center) http://www.ensembl.org/
- AVID alignment program

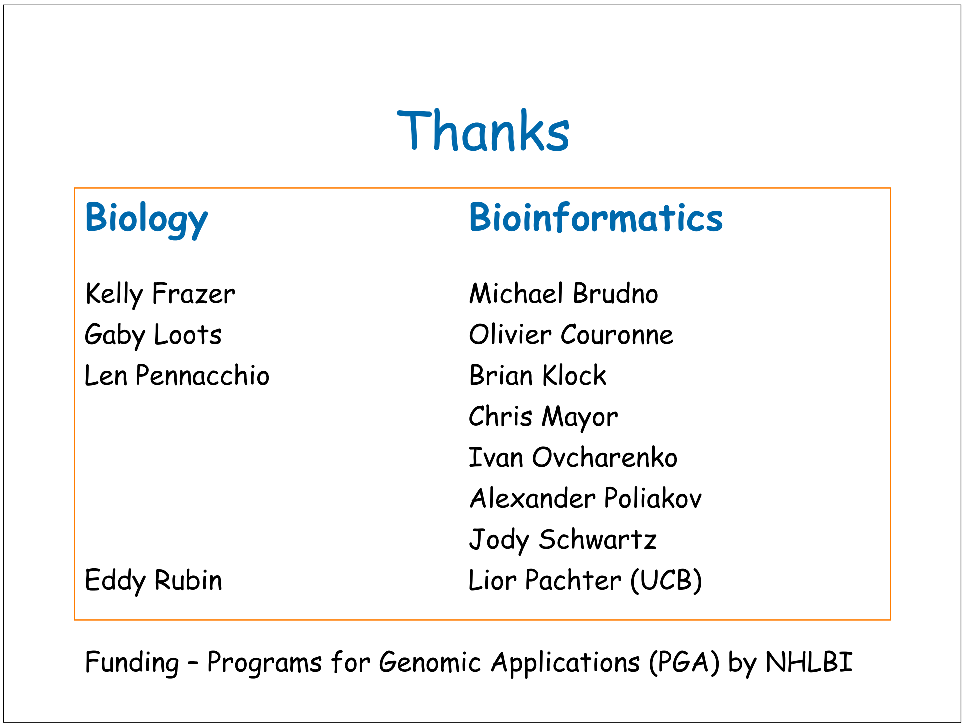
  http://baboon.math.berkeley.edu/~syntenic/avid.html
- SLAM comparative gene prediction program
  http://bio.math.berkeley.edu/slam/mouse/
- PSU group's MHC Human-Mouse comparison results
  http://bio.cse.psu.edu/mousegroup/MHC/
- PSU Pipmaker program http://bio.cse.psu.edu/pipmaker/

Towards Better VISTAs

Information from a Single Sequence Alone



Multi-Organism High Quality Sequences



Thanks

Biology	Bioinformatics
Kelly Frazer	Michael Brudno
Gaby Loots	Olivier Couronne
Len Pennacchio	Brian Klock
	Chris Mayor
	Ivan Ovcharenko
	Alexander Poliakov
	Jody Schwartz
Eddy Rubin	Lior Pachter (UCB)

Funding – Programs for Genomic Applications (PGA) by NHLBI